Clim Dyn (2014) 43:671–688
DOI 10.1007/s00382-013-1922-6

Climate model benchmarking with glacial and mid-Holocene climates

S. P. Harrison · P. J. Bartlein · S. Brewer · I. C. Prentice · M. Boyd · I. Hessler · K. Holmgren · K. Izumi · K. Willis

Received: 26 March 2013 / Accepted: 15 August 2013 / Published online: 30 August 2013

Abstract Past climates provide a test of models' ability to predict climate change. We present a comprehensive evaluation of state-of-the-art models against Last Glacial Maximum and mid-Holocene climates, using reconstructions of land and ocean climates and simulations from the Palaeoclimate Modelling and Coupled Modelling Intercomparison Projects. Newer models do not perform better than earlier versions despite higher resolution and complexity. Differences in climate sensitivity only weakly account for differences in model performance. In the glacial, models consistently underestimate land cooling (especially in winter) and overestimate ocean surface cooling (especially in the tropics). In the mid-Holocene, models generally underestimate the precipitation increase in the northern monsoon regions, and overestimate summer warming in central Eurasia. Models generally capture large-scale gradients of climate change but have more limited ability to reproduce spatial patterns. Despite these common biases, some models perform better than others.

Keywords Climate-model evaluation · Climate sensitivity · Last Glacial Maximum · Mid-Holocene monsoons · Palaeoclimate Modelling Intercomparison Project

1 Introduction

Simulations of the Last Glacial Maximum (LGM, ca 21,000 years ago) and mid-Holocene (MH, ca 6,000 years ago) are included for the first time in the set of climate-model simulations performed by the Coupled Model Intercomparison Project (CMIP5: Taylor et al. 2012). This development recognizes the unique opportunity to use paleoclimates to evaluate, or benchmark, the models that are used for future climate projections (Taylor et al. 2011;

Electronic supplementary material The online version of this article (doi:10.1007/s00382-013-1922-6) contains supplementary material, which is available to authorized users.

S. P. Harrison (✉) · I. C. Prentice · I. Hessler · K. Willis
Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia
e-mail: sandy.harrison@mq.edu.au

S. P. Harrison
Centre for Past Climate Change and School of Archaeology, Geography and Environmental Sciences, University of Reading, Whiteknights, Reading RG6 6AH, UK

P. J. Bartlein · K. Izumi
Department of Geography, University of Oregon, Eugene, OR, USA

S. Brewer
Geography Department, University of Utah, Salt Lake City, UT, USA

I. C. Prentice
AXA Chair of Biosphere and Climate Impacts, Department of Life Sciences and Grantham Institute for Climate Change, Imperial College, Silwood Park, Ascot SL5 7PY, UK

M. Boyd · K. Holmgren
Bert Bolin Centre for Climate Research, Stockholm University, 106 91 Stockholm, Sweden

M. Boyd · K. Holmgren
Department of Physical Geography and Quaternary Geology, Stockholm University, 106 91 Stockholm, Sweden

I. Hessler
MARUM, Centre for Marine Environmental Sciences, University of Bremen, Bremen, Germany

Braconnot et al. 2011). The LGM and MH are geologically recent times with strong and contrasting climate forcing (Braconnot et al. 2012), the response to which is documented by abundant palaeoenvironmental records (Harrison and Bartlein 2012). Braconnot et al. (2012) showed that climate models can successfully reproduce the first-order patterns of past climate changes, including the land-sea contrast and high-latitude amplification of temperature change and the impacts of changes in monsoon circulation on precipitation patterns, but are unable to reproduce the magnitude of observed regional changes in climate. Quantitative evaluation of climate models using paleoclimates has generally focused on specific regions or phenomena (e.g. Joussaume et al. 1999; Otto-Bliesner et al. 2009; Zhang et al. 2010; Valdes 2011; Braconnot et al. 2012; Harrison and Bartlein 2012). More systematic testing is now possible using global syntheses of paleoclimate reconstructions (e.g. MARGO Project Members 2009; Leduc et al. 2010; Bartlein et al. 2011; Schmittner et al. 2011).

Here we present an evaluation of the MH and LGM simulations from the CMIP5 archive against ten data sets that include annual and seasonal climate variables over the land and oceans. We begin by comparing the global performance and the geographic expression of simulated climate changes in the CMIP5 simulations with an earlier generation of LGM and MH simulations made during the second phase of the Palaeoclimate Modelling Intercomparison Project (PMIP2), to determine whether recent improvements in modelling schemes or the use of higher-resolution models has resulted in differences in model performance. We then provide an evaluation of the CMIP5 and PMIP2 models against the observational benchmarks, using standard metrics to assess different aspects of the goodness-of-fit and biases of individual models. Finally, we present an overall assessment of model performance to address the question of whether some models consistently perform better than others.

2 Data and methods

2.1 The benchmark data sets

There has been a long history of making quantitative climate reconstructions using biological, isotopic and geochemical records from land and ocean sites. More recently, community efforts have focused on creating synthetic global data sets for specific times and types of record. We have created benchmark data sets through combining the unique data points from existing syntheses, specifically pollen- and plant macrofossil-based reconstructions of land climate at the LGM and MH from Bartlein et al. (2011), surface ocean reconstructions from the MARGO data set for the LGM (MARGO Project Members 2009) and from the GHOST data set for the MH (Leduc et al. 2010), additional LGM land (5) and ocean (25) records from Schmittner et al. (2011) and Antarctic ice-core estimates of LGM temperatures from Braconnot et al. (2012). (Details of the original data sets, including their treatment of reconstruction uncertainties, are given in the SI.)

In addition to combining existing syntheses, we have compiled and evaluated information from available speleothem records. Analyses of the stable isotopes, trace elements, luminescence and fabric of calcium carbonate precipitates (speleothems) from limestone caves can provide precisely dated information about long-term climate variability (e.g. Fairchild et al. 2006; Lachniet 2009). However, the interpretation of these records, and the unambiguous derivation of quantitative climate reconstructions, is dependent on site-specific conditions. We have reviewed and evaluated the published speleothem records based on the reliability of the age model, the presence of samples dated to either the MH or LGM, the availability of information about modern-day conditions (thus allowing quantitative calibration of the records), and the robustness of the inferences about the palaeorecord (see SI for details of the screening procedure). Of the 65 records examined, lack of information about modern conditions means that climate anomalies can only be provided for 37 sites (Figure S1) and only 6 of these provide quantitative reconstructions (see SI). Nevertheless, these speleothem reconstructions are useful because they provide information from regions not covered by other kinds of data.

Site-based reconstructions from each of these sources were combined to produce new quantitative reconstructions (with uncertainty estimates) of 10 climate variables on a common 2° by 2° resolution land or ocean grid. The climate variables are mean annual temperature (MAT), mean temperature of the coldest month (MTCO), mean temperature of the warmest month (MTWA), accumulated temperature sum during the growing season (GDD5), mean annual precipitation (MAP), and the ratio (α) of actual to equilibrium evapotranspiration (i.e. evapotranspiration from a large, homogeneous well-watered surface: Raupach 2001) over land (Figures S2 and S3); summer (SSTsum), winter (SSTwin) and annual (SSTann) sea-surface temperatures; and number of months with sea-ice cover (SInmon) over the ocean (Figures S4 and S5). The gridded estimates were produced using simple averaging and the grid-cell uncertainty was calculated as the pooled estimate of the standard error. Inspection of these values shows that the uncertainties are much smaller than the standard errors of spatial averages of the reconstructions. These uncertainties are explicitly taken into account in the calculation of fuzzy distance (see SI).

2.2 Climate model simulations and processing of model outputs

The LGM and MH simulations are equilibrium experiments. The insolation, ice sheet and GHG forcings used for each experiment are described in Braconnot et al. (2012). Although other forcings are the same between the PMIP2 and CMIP5 experiments, the LGM ice-sheet forcing is different: in PMIP2 the ice sheets were specified from Peltier (2004) while in CMIP5 a blended product made from three more recent ice-sheet reconstructions was used (Braconnot et al. 2012).

We use the LGM and MH simulations in the CMIP5 (http://cmip-pcmdi.llnl.gov/cmip5/data_portal.html) and PMIP2 (http://pmip2.lsce.ipsl.fr/database/access/opendap.shtml) archives as of 15th August 2012 (Table S6). Most of these simulations are made with ocean–atmosphere (OA) models. Some of the PMIP2 models simulated vegetation dynamics explicitly (i.e. were fully-coupled ocean–atmosphere-vegetation models, OAVs). Processes associated with the terrestrial and marine carbon cycle were ignored in PMIP2 experiments, but are included as interactive components of some of the models (here designated as OACs) in CMIP5. MH simulations are available for 13 OAs and 6 OAVs from the PMIP2 archive and 10 OA simulations and 5 OAC simulations from the CMIP5 archive. LGM simulations are available for 7 OA simulations and 2 OAV simulation from the PMIP2 archive, and 3 OA and 3 OAC simulations from the CMIP5 archive. There was no MH SSTann data archived by August 15th 2012 for the ECHAM, EARTH, FGOALSG2, and FGOALS2 models, no LGM sea-ice data for the IPSL4 model, and no SST or sea-ice data for the COSMOS model. For these models, comparisons were restricted to the subset of variables available.

Long-term means were calculated from the archived time-series data on individual model grids for five climate variables: near-surface air temperature (*tas*), precipitation flux (*pr*), cloud-area fraction (*clt*), sea-surface temperature (*tos*), and sea-ice fraction (*sic*). The temperature, precipitation and cloud-cover means were bi-linearly interpolated to a common 0.5° grid, in order to calculate bioclimatic variables (GDD5, MTWA, MTCO, MAT, MAP and α) for comparison with the benchmark data sets. Bioclimatic variables were calculated using the anomalies on the 0.5° grid using the approach of Prentice et al. (1992). The original routines of Cramer and Prentice (1988) and Prentice et al. (1993) were modified to include snow-moisture accounting and to use a multi-layer soil-characteristic data set (IGBP-DIS). Finally, the bioclimatic variables and sea-surface temperature and sea-ice fraction data were then regridded to the 2° × 2° grid of the palaeo-reconstructions, using simple averaging, to facilitate comparisons and for the calculation of ensemble-averages of model output. A detailed description of the model output processing is given in the SI.

2.3 Comparison of PMIP2 and CMIP5 simulations

To compare the two generations of simulations, we calculated ensemble averages of the climate variables. The differences between the ensemble-average anomalies for individual variables illustrate the change in simulated patterns between the two generations (PMIP2 and CMIP5) of simulations. Given the number of available variables and grid cells (16,200), such comparisons will inevitably reveal many large or "significant" differences between the ensemble averages for individual variables. However, the issue is whether the different generations of simulations differ overall.

To more formally assess the differences in the ensemble averages between generations of simulations, we calculated Hotelling's T^2 statistic (Wilks 2011) on the climate anomalies, for each of the 2° × 2° grid cells for particular combinations of variables. Hotelling's T^2 is a multivariate generalization of the ordinary t-statistic that is appropriately used to examine differences in climate-model simulations (Chervin and Schneider 1976). Like the *t*-statistic, Hotelling's T^2 scales the difference between the means by a measure of the variability of the groups of observations being compared such that small values of the statistic could result from small differences in the means, or large variability among (in this case) models. The climate anomalies are approximately normally distributed and have similar variance between the two sets of simulations. Hotelling's T^2 is known to be sensitive to the trade-off between the number of observations and the number of variables (Rencher 2002), and so we limited the number of variables considered. In its application here, the multiple variables include either bioclimatic variables that are available globally (i.e. MAT, MTCO, MTWA and MAP), or selections of monthly temperature and precipitation (*tas* and *pre* for January, April, July and October), the observations are the individual model simulations grouped by simulation generation, and the null hypothesis is that the ensemble means are equal between groups. Separate comparisons were made for the MH and LGM simulations. Comparison of subsets of simulations within generations (e.g. CMIP5 OA vs. CMIP5 OAC) is not warranted owing to sample-size considerations.

A test statistic and associated significance level (*p* value) is obtained for each grid point for a specific comparison, and it is likely that some number of these local tests will appear to be significant (i.e. $p < 0.05$) simply by chance, and so a simple count of those tests to determine a global "field significance" may be misleading (the "false

discovery rate" issue, see Wilks 2006). In other words, in the 16,200 individual tests, we should expect that five percent (810) would appear as significant even if the null hypothesis of no difference between simulation generations were true. We therefore applied the approach of Ventura et al. (2004) to evaluate the number of "significant" hypothesis tests in each comparison. In this approach, the sorted individual "local" p values (one at each grid point) are compared with a progression of false discovery rate (FDR) criteria (see Ventura et al. 2004, eqn. 2), and the proportion of the local p values that do not exceed those criteria provides support (or lack thereof) for rejecting a global null hypothesis of no difference between simulation generations. In practice, the FDR approach requires a larger number (than five percent) of local tests to have p values below the usual threshold (i.e. $p < 0.05$) before declaring the overall hypothesis of no difference in anomalies to be false. The anomaly patterns being compared are generally large in spatial scale, leading to correlations among the local tests, but Wilks (2006) shows that the FDR procedure is still robust in such a situation.

2.4 Metrics for comparison of reconstructed and simulated climate variables

Many metrics, each with different properties, have been used in the geosciences literature to compare observed and modelled quantities. Rather than focus on a single metric of model skill, we use a range of different measures to examine different aspects of model performance. We use medians and the interquartile range (IQR), calculated using only those grid cells where there are observations, to provide a basic measure of global agreement between model and observation. The IQR provides a measure of spatial variability in climate anomalies; comparison of simulated and reconstructed IQR therefore assesses the agreement in the amplitude of the anomalies. We assess the similarity of simulated and observed geographic patterning in climate anomalies using Kendall's rank correlation coefficient tau (τ), which measures the similarity or difference of spatial patterns regardless of magnitudes (Kendall 1938). We present values as $1 - \tau$, which takes the values of 0 when patterns are identical and 1 when there is no correlation between them. Distance measures provide an overall measure of similarity. Fuzzy distance is a measure of the distance or dissimilarity between two quantities which takes account of measurement uncertainties: the effect of increasing uncertainty is to increase the fuzzy distance (Guiot et al. 1999; Tran and Duckstein 2002). The mathematical description of each of these metrics is given in the SI.

These four metrics are calculated for each of the individual model simulations from PMIP2 and CMIP5. In addition, we calculate the metrics for various subsets of the models (all models, all the PMIP2 simulations, all the PMIP2 OA simulations, all the PMIP2 OAV simulations, all the CMIP5 simulations, all the CMIP5 OA simulations and all the CMIP5 OAC simulations). For each subset, we create an ensemble average by calculating average values for each grid cell across the suite of models; the global metrics are then calculated from these ensemble averages. For example, the median bias of the subset of CMIP5 models is calculated from an ensemble model created by taking the median value of the grid-cell values of all the CMIP5 models, grid cell by grid cell. The global metric is the "multi-model ensemble median bias" (see Gleckler et al. 2008).

2.5 Metric for overall evaluation of model performance

Rather than devising a single "skill score" for overall performance, which necessarily involves making arbitrary choices about the relative importance of individual variables and types of bias, we evaluate model performance for each climate variable and metric. This also obviates overinflation of the skill score because of partial correlations among the variables. Following Gleckler et al. (2008) the metrics are normalized by the median model error to yield an evaluation of how well a given model compares to the typical model error. The median model error is calculated as the median of the global error for each individual model from CMIP5 and PMIP2. Thus, the normalization procedure uses the "median errors within the distribution of individual model errors" not the "multi-model ensemble median error". This allows the metrics of the ensemble models created from the various subsets of models described above to be compared with the median model error. The normalization procedure yields negative values for models that perform better than the median model and positive values for models that perform worse. Values <-0.5 indicate that the models are 50 % better than the median model error, whereas values >0.5 indicate models that are 50 % worse than the median model error. In order to visualize these results, the models are ordered from best to worst, either based on an average of the normalized values across all of the variables and metrics, or alternatively for a single metric across all the variables.

3 Results

3.1 Comparison of CMIP5 and PMIP2 simulations

The CMIP5 palaeo-simulations were made with the version of each model that is used for future projections, and at the same resolution. Many PMIP2 simulations were made with

lower-resolution and/or older model versions than those now used for future projections. However, the range of changes in seasonal temperature and precipitation, both globally and regionally, is similar for both groups of models (Fig. 1). Coupled ocean–atmosphere-vegetation (OAV) models tend to show larger changes in climate (e.g. increased MH warming over land, increased LGM cooling) than the OA models; but similar changes are produced by some models without considering vegetation feedbacks, and the OA versions of these OAV models already tend to show stronger changes than other OA models. In contrast, there is no systematic difference between the mean climates simulated by the CMIP5 OA and OAC models.

The spatial patterns of simulated climate change in the two sets of simulations are broadly comparable (Figs. 2, 3). There are some systematic differences in the anomalies related to the differences in the specification of the LGM ice sheets in the two generations of simulations (right-hand column of Fig. 2), including higher temperatures over the Laurentide ice sheet in the CMIP5 simulations, and generally lower temperatures in the northern mid-latitudes (Braconnot et al. 2012). The southern oceans in the CMIP5 simulations are somewhat warmer than in the PMIP2 simulations. The largest differences in the anomaly patterns for temperature in the MH simulations lie over northern North America, where the CMIP5 anomalies are a little lower than the PMIP2 anomalies. The differences in the LGM anomaly patterns of precipitation to some extent reflect the temperature anomaly differences, and additionally show some dipole patterns in the tropics reflecting the latitudinal movement of the intertropical convergence zone. The differences in precipitation anomalies for the MH (Fig. 3) are generally smaller in magnitude than those for the LGM, and similarly show some latitudinal dipoles for precipitation in the tropics.

In general, the patterns of "significant" tests (i.e. $p < 0.05$) obtained from the local Hotelling's T^2 statistics are quite noisy (Fig. 4), and there is little relation between the p values and the patterns of the largest anomaly differences. For the tests involving MAT, MTCO, MTWA and MAP in the MH simulations there is a relatively large area of p values <0.05 over northern North America, and some latitudinally organized patterns in the tropics, but the total number of p values <0.05 is still relatively small (1,668 out of 16,200), and none of the p values fall below the individual FDR threshold values (i.e. there are no more "significant" p values than would be expected by chance). For the comparisons involving *tas* and *pre* in the MH simulations, the number of p values <0.05 is larger (2,293), but again none of the individual ranked p values fall below the FDR threshold values. For the LGM simulations, the numbers of p values less than 0.05 are smaller than for the MH simulations (729 for the comparisons involving MAT, MTCO, MTWA and MAP, and 879 for the comparisons involving *tas* and *pre*). Consequently neither the map patterns of the local Hotelling's T^2 statistics, nor the number of "significant" local tests, provide support for the idea that the two generations of simulations differ from one another. Thus, the analysis provides no evidence that the CMIP5 and PMIP2 simulations differ systematically.

3.2 Evaluation of LGM simulations

3.2.1 The glacial ocean

The ocean temperature, over the regions for which there are SST reconstructions, was 1.9 °C colder at the LGM. Year-round cooling is consistent with the year-round forcing caused by the presence of large northern hemisphere ice sheets and lowered greenhouse gas concentrations. Most models overestimate the ocean cooling (Fig. 5). Two models (CCSM, MIROC) produce good estimates of the median change in annual sea-surface temperature (i.e. within 0.1 °C of the reconstructed median value). In five cases the median bias is larger than 0.5 °C (Table S9) with the most extreme biases (>0.8 °C) shown by the HadCM3 (OAV) and ECHAM (OAV) models. Comparison of the ensemble averages (Table S7) shows that the OAV simulations are more inconsistent with the reconstructions than the PMIP2 OA or CMIP5 models.

According to the reconstructions, ocean cooling occurs equally in both seasons. The models show larger cooling in summer (Fig. 5, Table S9). The PMIP2 OAV simulations produce colder oceans in both seasons than the PMIP2 OA or CMIP5 models, and again are more inconsistent with the seasonal reconstructions (Table S7). The mismatch between simulated and reconstructed annual (and seasonal) SSTs arises because the models overestimate the cooling in the tropics (30°N–30°S) and northern high-latitudes (>75°N). Conversely, they underestimate the cooling in the mid-latitudes (Figure S6). Between 45°–60°N, the ensemble median bias is larger than 1 °C. Some models have a bias >2° in this region (i.e. the simulated cooling is only about half of the reconstructed cooling).

The inter-quartile range (IQR) provides a measure of the spatial heterogeneity of an observed and/or simulated climate variable. This is not a measure of uncertainty of the median value, but rather shows the degree to which there is geographic variability in a given quantity. The IQR of the reconstructed SST anomalies is large (Fig. 5), both globally and in any zonal belt. Models consistently underestimate this variability (i.e. they do not capture the heterogeneity seen in the reconstructions even when gridded to the same scale as the model outputs) except north of 60°N. Globally, the IQR of the models is between 18 and 72 % of the reconstructed IQR of 2.7 °C, which suggests

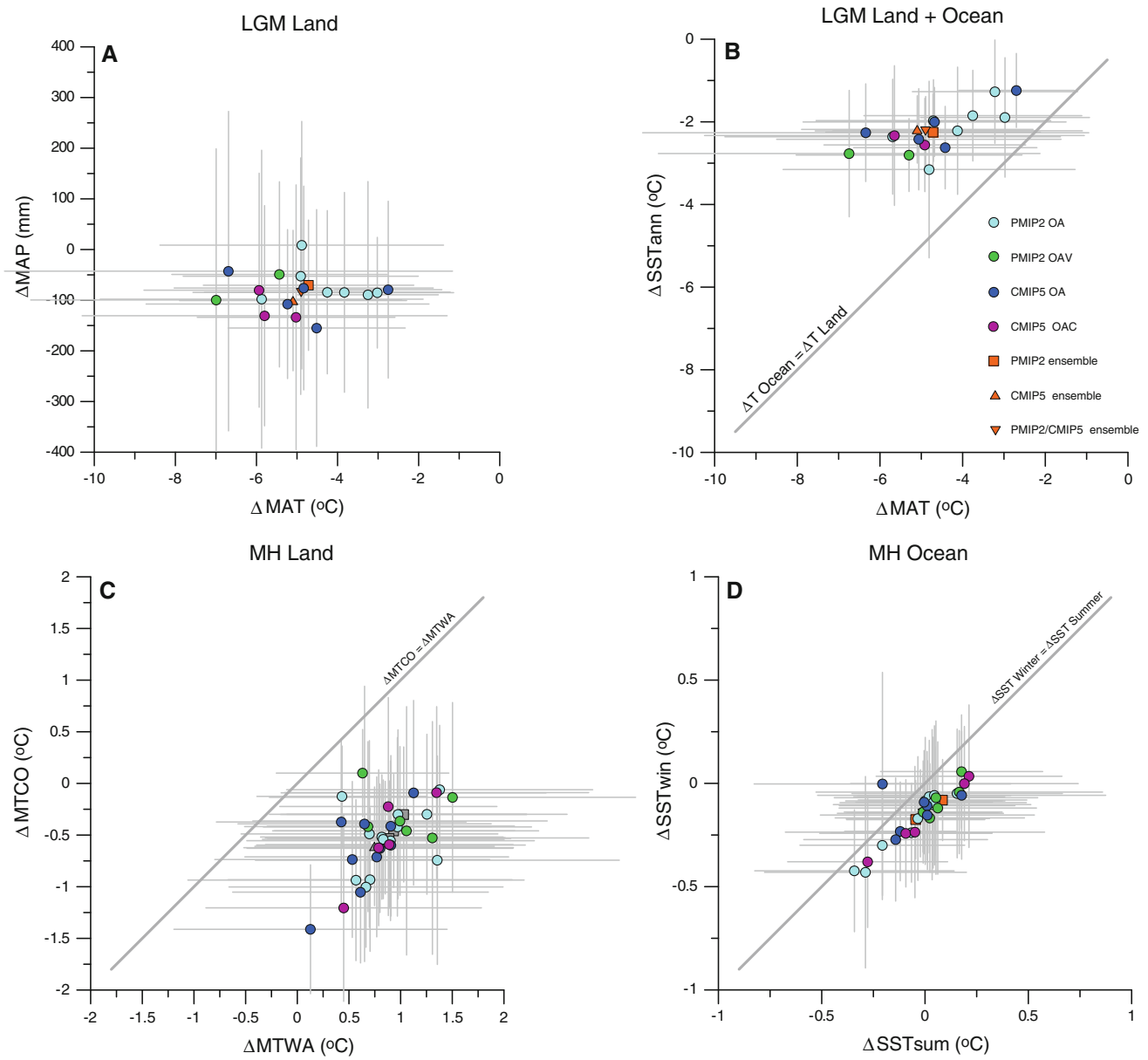


Fig. 1 Anomalies in global climate at the Last Glacial Maximum (LGM, ca 21,000 year BP) and the mid-Holocene (MH, ca 6,000 year BP) as simulated by the individual PMIP2 and CMIP5 models compared to the ensemble average. For the LGM, climate space is defined by (**a**) the change in mean annual temperature (MAT) and mean annual precipitation (MAP) because the changes in climate forcing operate to produce year-round cooling and drying compared to present day. We also show (**b**) the relationship between changes in annual sea-surface temperature (SSTann) and MAT. This graph shows the expected enhancement of temperature changes over land compared to the ocean. The change in insolation forcing during the MH produces primarily seasonal responses, so the climate space during this interval is defined by (**c**) the changes in mean temperature of the coldest month (MTCO) and mean temperature of the warmest month (MTWA). We also compare (**d**) seasonal changes in ocean temperatures in summer (SSTsum, June, July, August in the northern hemisphere and January, February, March in the southern hemisphere) and winter (SSTwin, January, February, March in the northern hemisphere and June, July, August in the southern hemisphere)

that the spatial correlation scale in the models may be longer than seen in the reconstructions.

There is considerable geographic patterning in reconstructed climate changes over land and ocean at the LGM. Cooling is most pronounced close to and downwind of the northern hemisphere ice sheets, less marked upwind of the Laurentide ice sheet, and small in the tropics. The spatial patterns in reconstructed LGM SSTann anomalies, as measured by $1 - \tau$, are not well predicted by the models (Table S9), with values ranging from 0.72 (ECBILT) to 0.96 (CNRM5). The seasonal SST anomalies show no correlation with the reconstructions, with all of the models

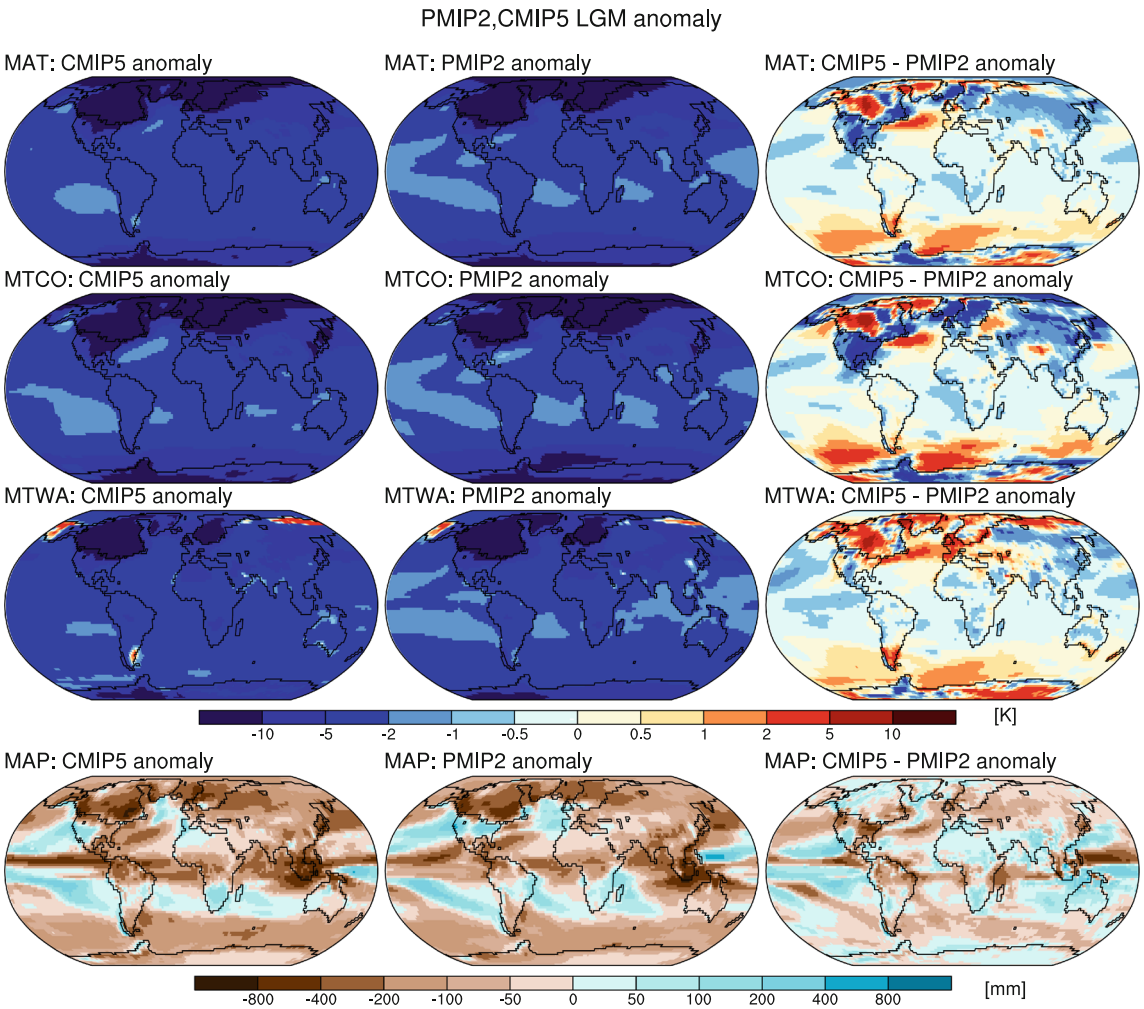


Fig. 2 Simulated changes (anomalies between the experiment and the pre-industrial control) in mean annual temperature (MAT), mean temperature of the coldest month (MTCO), mean temperature of the warmest month (MTWA) and mean annual precipitation (MAP) at the Last Glacial Maximum (LGM) for the CMIP5 ensemble (*left hand column*) and the PMIP2 ensemble (*centre column*). The ocean temperatures are sea-surface temperature, except over areas with sea ice where air temperature is used (see SI). The difference between the two sets of simulations is also shown (*right hand column*)

obtaining values close to 1 for both SSTsum (0.94–1.03) and SSTwin (0.95–1.02).

The fuzzy distance provides an overall measure of model performance. Models with different median biases, for example, can nevertheless have a similar distance score if the model with the larger bias captures the spatial variability or patterning better. For example, GISS.E2 has a slightly better overall score than MPI (ESM) for SSTann (Table S9) because although GISS.E2 displays a larger bias it has a more realistic range of variability. The distance measures confirm that models generally reproduce SSTann (1.18–1.50) better than SSTwin (1.28–1.71), which in turn is better than SSTsum (1.44–1.93), consistent with the unrealistically larger cooling in summer than winter. The distance measures for SInmon show that the mismatch between simulated and observed sea-ice cover is typically one to 2 months, although FGOALS1 has a bias of 4 months.

3.2.2 *The glacial continents*

The reconstructions show year-round cooling over the continents at the LGM (Fig. 5). Based on regions with reconstructions, MAT was reduced by 6.4 °C. Winter cooling was greater than summer cooling (−9.6 °C compared to −4.3 °C). All but two models underestimate the reconstructed annual cooling, with the largest median bias nearly 3.5 °C and eight models having a bias larger than 1 °C. MIROC (ESM) and HADCM3 (OAV) overestimate the reconstructed cooling (Fig. 5). The OA version of HadCM3 also produces a greater year-round cooling than most other models (close to the reconstructed change), but

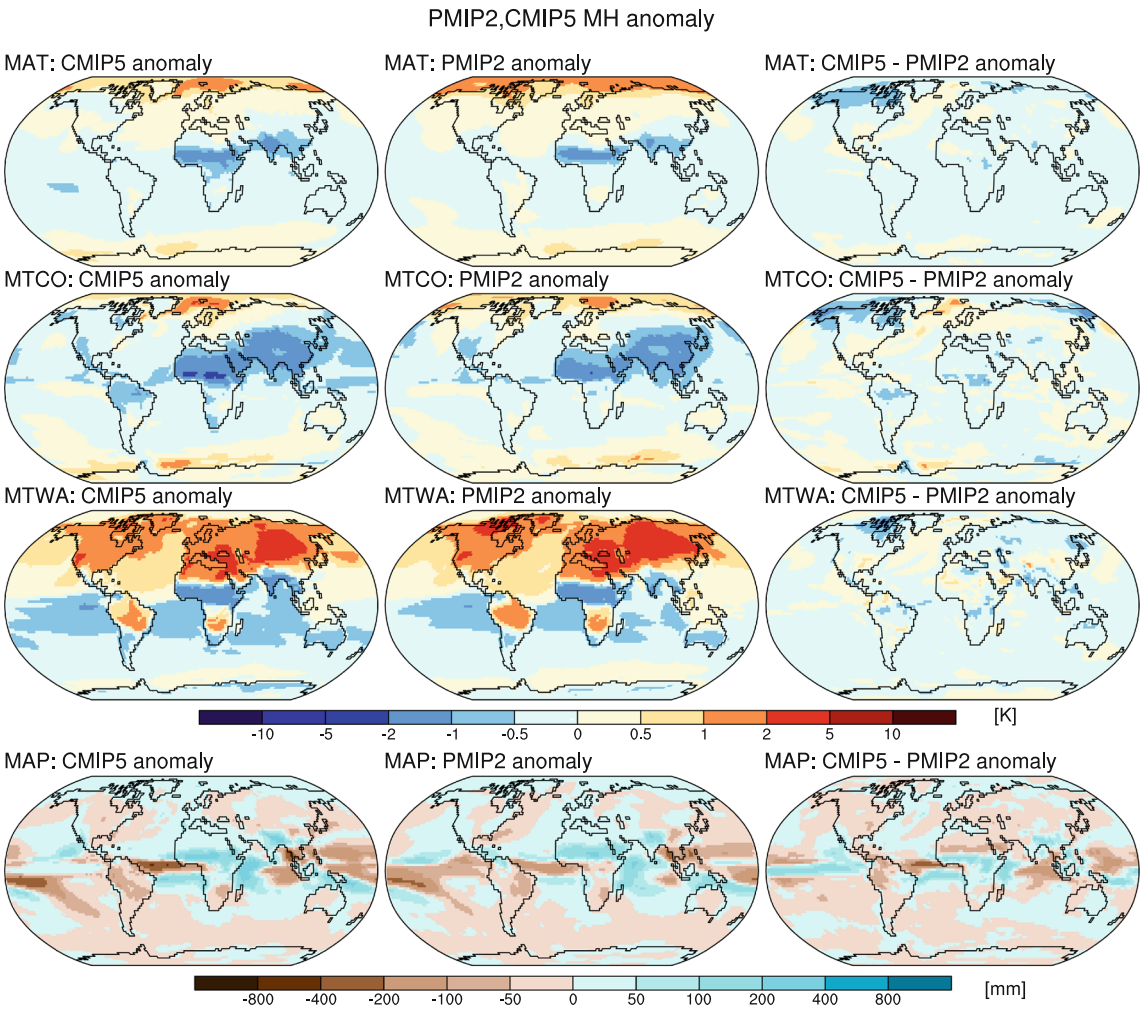


Fig. 3 Simulated changes (anomalies between the experiment and the pre-industrial control) in mean annual temperature (MAT), mean temperature of the coldest month (MTCO), mean temperature of the warmest month (MTWA) and mean annual precipitation (MAP) at the mid-Holocene (MH) for the CMIP5 ensemble (*left hand column*) and the PMIP2 ensemble (*centre column*). The difference between the two sets of simulations is also shown (*right hand column*)

the OA version of MIROC is comparatively warm compared to other models. The underestimation of MAT is driven by underestimation of winter cooling. All the models underestimate the MTCO reduction; the smallest median bias is +2.4 °C and the largest +7.3 °C. Some models underestimate and some overestimate the reconstructed summer cooling. All models underestimate the LGM reduction in MAP over land, consistent with the underestimation of the change in temperature (Li et al. in press). As a result, most models also underestimate the increase in aridity (reduction in α). The bias in median MAP ranges from +10 to +334 mm; nine models produce changes in median MAP that are less than half of the reconstructed change. The discrepancies between simulated and reconstructed α are smaller, because the smaller-than-reconstructed reduction in precipitation is offset by the smaller-than-reconstructed reduction in temperature.

The IQR of modeled LGM land climates is smaller than reconstructed for most, but not all, variables (Fig. 5). The simulated spatial variability in MTWA is consistently smaller than shown by the reconstructions: simulated MTWA IQR is 29–86 % of the reconstructions. With the exception of the COSMOS model (112 %), the simulated IQR of GDD5 is between 40 and 70 % of the reconstructions. However, the simulated IQR of MAT ranges from much smaller to somewhat larger (27–145 %), that for MTCO is from 26 to 103 %, and that for MAP is 53–116 %. The variability in α is always larger than shown by the reconstructions (155–419 %).

The geographic patterns in the sign of the changes over land at the LGM during winter (MTCO) are in general moderately well predicted, with values of $1 - \tau$ ranging from 0.58 to 0.82 (Table S9). The prediction of GDD5 is also moderately good (0.58–0.82). Three models score 1 or

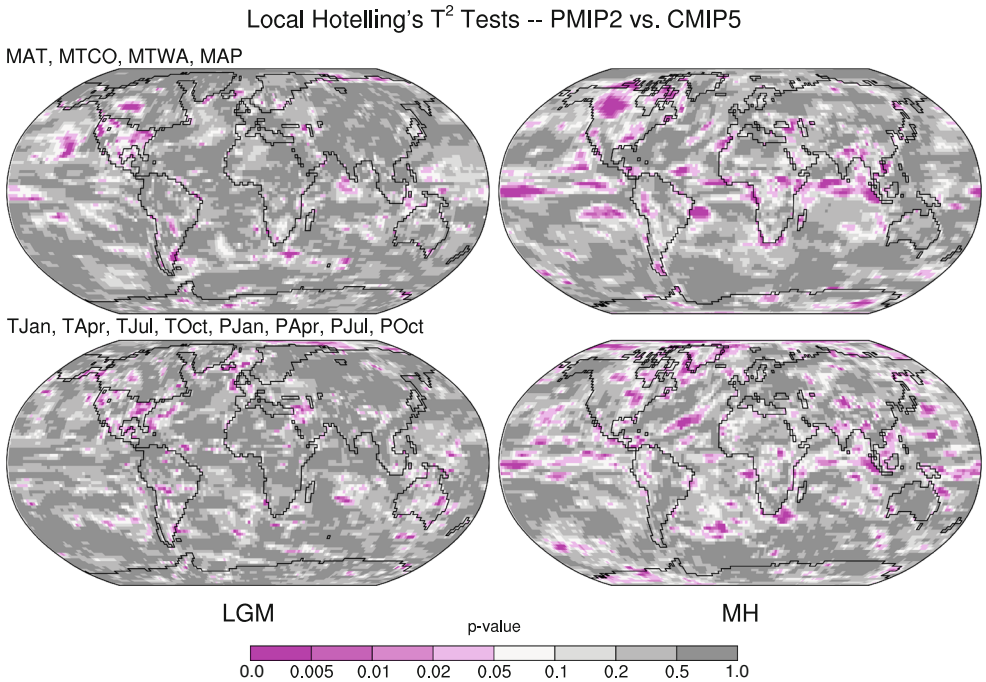


Fig. 4 Maps of the *p* values of Hotelling's T^2 for comparisons of the PMIP2 and CMIP5 ensembles. The *upper plots* show the results of tests using mean annual temperature (MAT), mean temperature of the coldest month (MTCO), mean temperature of the warmest month (MTWA) and mean annual precipitation (MAP) at the Last Glacial Maximum (LGM: *upper left hand plot*) and the mid-Holocene (MH: *upper right hand plot*). The *lower plots* show the results of tests using mean January temperature (Tjan), mean April temperature (Tapr), mean July temperature (Tjul) and mean October temperature (Toct) and mean precipitation for the same four months (Pjan, Papr, Pjul, Poct) for the LGM (*lower left hand plot*) and MH (*lower right hand plot*) respectively

>1 for MTWA, i.e. there is no correlation between the simulated and observed patterns, but the range for the other models (0.63–0.97) is comparable to the other seasonal variables. The simulation of the geographic patterns in MAT is also moderately good (0.57–0.89), with the exception of a single model (CNRM3.3), which scores 1. The simulation of geographic patterning in α is poor, with a range of values between 0.72 and 0.88 (Table S9). However, the simulation of the spatial pattern in precipitation is poorer, with nine models having values of $1 - \tau$ greater than 1. In general, the prediction of LGM temperature anomaly patterns appears to be better over land than over the ocean. This is apparent in comparing e.g. MAT with SSTann where 15 out of 17 models have $1 - \tau$ values of <0.80 for MAT compared to only 8 out of 16 models with values <0.80 for SSTann. However, the simulation of seasonal climate over land is very much better than seasonal climates over the ocean, where all of the models have values of close to 1 (i.e. no correlation) for geographic patterning over the ocean.

The fuzzy distance scores (Table S9) suggest that model performance is better for MAT (2.21–3.57) than for seasonal temperatures (MTCO: 3.61–6.24; MTWA 3.05–6.30). This is probably because, despite generally underestimating the annual cooling, the models capture the spatial variability of temperature changes moderately well. The range of the distance scores for MAP (135.91–387.37) and α (0.09–0.17) reflect the differences in the median biases: there is a large range for MAP but only small differences in the scores for α between the models.

3.3 Evaluation of mid-Holocene simulations

3.3.1 The mid-Holocene ocean

According to the reconstructions, the global ocean in the MH was slightly warmer than today (for regions with data); none of the models reproduce this (Fig. 6). The largest median biases are larger than 0.5 °C. OAV simulations produce warmer oceans than the OA simulations (Fig. 6, Tables S8 and S10), and are therefore more realistic. As at the LGM, the simulated cooling signal in SSTann is a reflection of cooling in the tropics (Figure S7). The models underestimate the reconstructed warming in northern mid-latitudes (30°–75°N). The models do not show warmer conditions in the southern mid-latitudes, which is inconsistent with the changes inferred from the limited number of reconstructions available. The seasonal nature of the insolation forcing leads to seasonal variations in SSTs (Fig. 3), with most models simulating a warmer ocean in

summer and colder conditions in winter than today in the northern hemisphere (Fig. 3). Individual site-based reconstructions suggest that summer warming and winter cooling are plausible (Yu et al. 2005; Morimoto et al. 2007; Giry et al. 2012).

The observed IQR of SSTann is 1.2 °C, considerably smaller than the estimate obtained for the LGM. However, as at the LGM, the models consistently underestimate the heterogeneity in SSTs, with IQR values between 17 and 56 % of the observed. The CMIP5 models generally show more heterogeneity at high northern and southern latitudes (Figure S10) than the PMIP models, but even so they underestimate the IQR values at these latitudes.

MH SST reconstructions are sparse (Figure S5), but show only moderate warming in the tropics and more pronounced warming in the northern mid- to high-latitudes. In the simulations (Fig. 3), the tropics are characterized by lower SSTs and there is a strong gradient in warming from the mid- to high-latitudes of the northern hemisphere. The geographic patterns in simulated MH SSTann anomalies, as measured by $1 - \tau$, are poorly predicted by the models (Table S11), with values ranging from 0.78 (MRI2) to 1.12 (HadCM3). Six models have values >1 (i.e. show some degree of anti-correlation with observations). The fuzzy distance measures (0.45–0.75) reflect this poor performance (Table S9).

3.3.2 *The mid-Holocene continents*

The MH, in regions with reconstructions, is characterized by slightly warmer summers, longer growing seasons, and increased precipitation relative to present (Fig. 6). This pattern reflects the distribution of the reconstructions, which are biased towards the northern hemisphere where insolation was increased in summer and reduced in winter relative to today (leading to little overall change in MAT) and increased seasonality amplified monsoonal rainfall. The models consistently underestimate the reconstructed change in MAP and α. As a result of underestimating the increase in precipitation (and α), and therefore presumably underestimating latent heat flux, the models overestimate summer warming by 0.56–2.27 °C. The simulated changes in MH winter temperature are not consistent between models. Some simulate warmer-than-present (and too warm) winters, others produce cooler-than-present (and too cool) winters. Biases in simulated MAT changes reflect the biases in the seasonal temperatures. Models that produce lower than reconstructed MAT tend to have more winter cooling and less summer warming than other models. Models that produce MAT warmer than reconstructed, tend to be warmer in both seasons.

Simulated MH land climates show consistently less spatial variability than the reconstructed climates. The IQR values of MAT (14–46 %), MTCO (14–56 %), MTWA (22–55 %), GDD (15–57 %) and MAP (17–39 %) from the simulations are consistently smaller than those of the reconstructions. The IQR of α ranges from 48 to 130 % of the reconstructions.

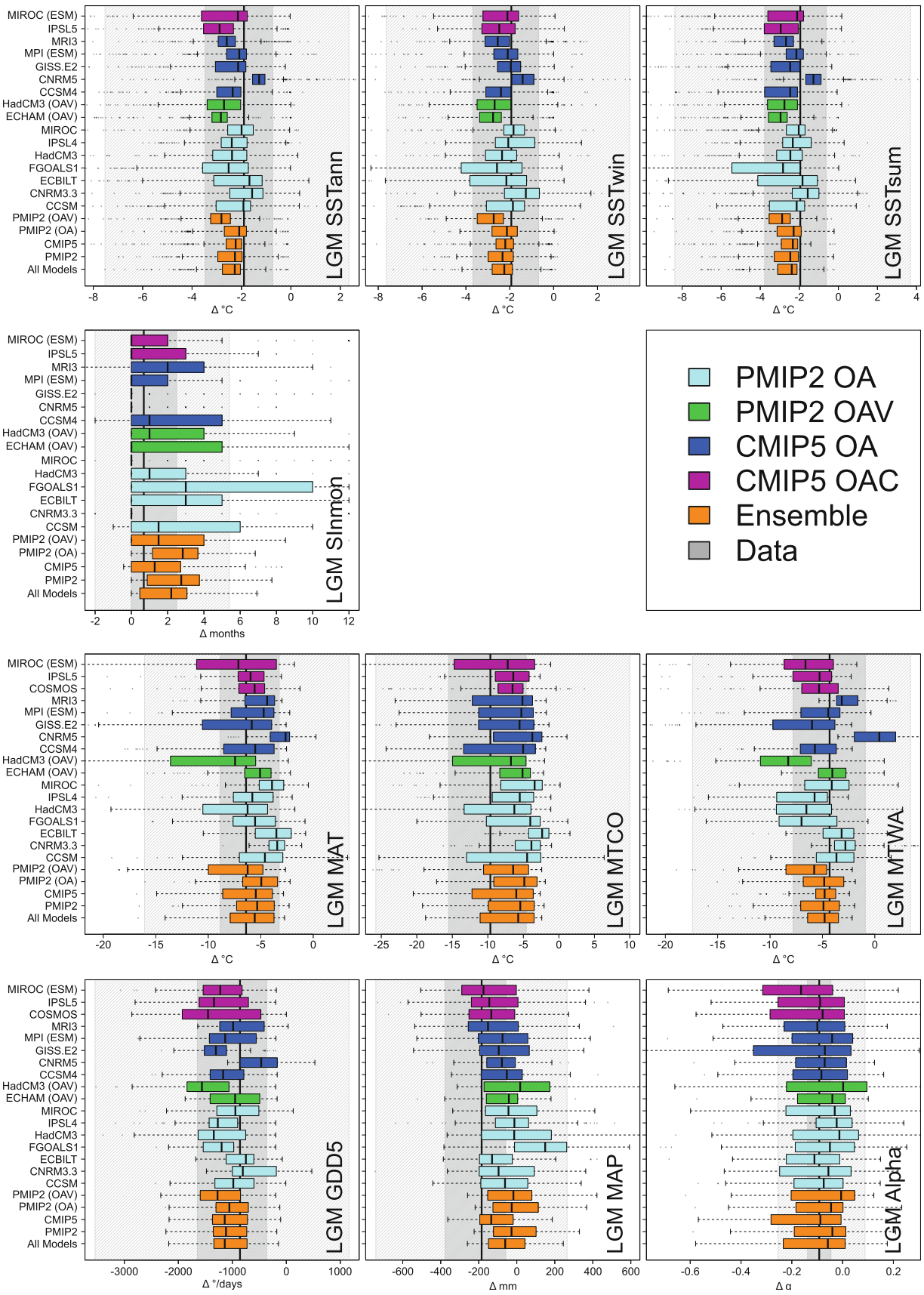
Fig. 5 Comparison of median and interquartile ranges (IQR) of observed and simulated climates at the LGM. The comparisons are made using only the model land (or ocean) grid cells where there are observations. Land climates are evaluated against the reconstructed bioclimatic variables (growing degree days above a threshold of 5 °C, GDD5; mean temperature of the warmest month, *MTWA;* mean temperature of the coldest month, *MTCO;* mean annual temperature, *MAT;* mean annual precipitation, *MAP;* the ratio of actual to equilibrium evapotranspiration, α). Ocean climates are evaluated against reconstructions of oceanic variables (summer sea-surface temperature, SSTsum; winter sea-surface temperature, SSTwin; mean annual sea-surface temperature SSTann; number of months with >40 % sea ice cover, SInmon). The median value of the observations is shown as a *black vertical line*, the IQR by *dark grey* shading and 5–95 percentile limits by *light grey* shading. The models are *color-coded* to show whether they are PMIP2 or CMIP5 simulations, and whether they are ocean–atmosphere (OA), ocean–atmosphere-vegetation (OAV) or OA carbon-cycle (OAC) models. The simulated median for each model is shown by a *vertical line*, the box represents the IQR and the whiskers the 5–95 percentile limits ▶

There is considerable geographic patterning in reconstructed climate changes over land at the MH (Figure S3), with a clear temperature gradient between the tropics and the northern extratropics. Enhanced northern-hemisphere monsoon circulation gives rise to increased precipitation in the monsoon core region but increased aridity in regions of descending air. Models' ability to simulate geographic patterning in the sign of climates in the MH is poorer than at the LGM. For example, $1 - \tau$ values for MAT range from 0.87 (IPSL and MRI2fa (OAV)) to 1.07 (FGOALS1). The best scores are in the range from 0.77 (for MAP) to 0.93 (for MTWA), and for all of the variables (other than MAT) between a third to a half of the models have scores of >1 indicating no correlation with the observations.

Despite the fact that the simulated geographic patterning is poorer in the MH than at the LGM, comparison of the fuzzy distance measures for each variable show that the overall performance of the models is better for the MH than the LGM (i.e. the biases are less extreme). Thus, the range of the fuzzy distances for MH MTCO is 1.34–1.91 compared to 3.61–6.24 for the LGM (Table S9, Table S10). The difference in the ranges is similar for MTWA (1.33–2.51 for the MH, 3.05–6.30 for the LGM). As in the LGM, the range for MAT (1.05–1.42) is better than for the seasonal temperatures. Similarly, the range of the fuzzy distances for MH MAP is 76–110, whereas the range for the LGM is 135–388.

3.4 Assessment of overall model performance

At the LGM (Fig. 7a, Table S8), most models perform only slightly better or worse (here defined as values between


MIROC (ESM)
IPSL5
MRI3
MPI (ESM)
GISS.E2
CNRM5
CCSM4
HadCM3 (OAV)
ECHAM (OAV)
MIROC
IPSL4
HadCM3
FGOALS1
ECBILT
CNRM3.3
CCSM
PMIP2 (OAV)
PMIP2 (OA)
CMIP5
PMIP2
All Models
LGM SSTann
LGM SSTwin
LGM SSTsum
Δ °C
LGM SInmon
Δ months
PMIP2 OA
PMIP2 OAV
CMIP5 OA
CMIP5 OAC
Ensemble
Data
COSMOS
LGM MAT
LGM MTCO
LGM MTWA
LGM GDD5
Δ °/days
LGM MAP
Δ mm
LGM Alpha
Δ α

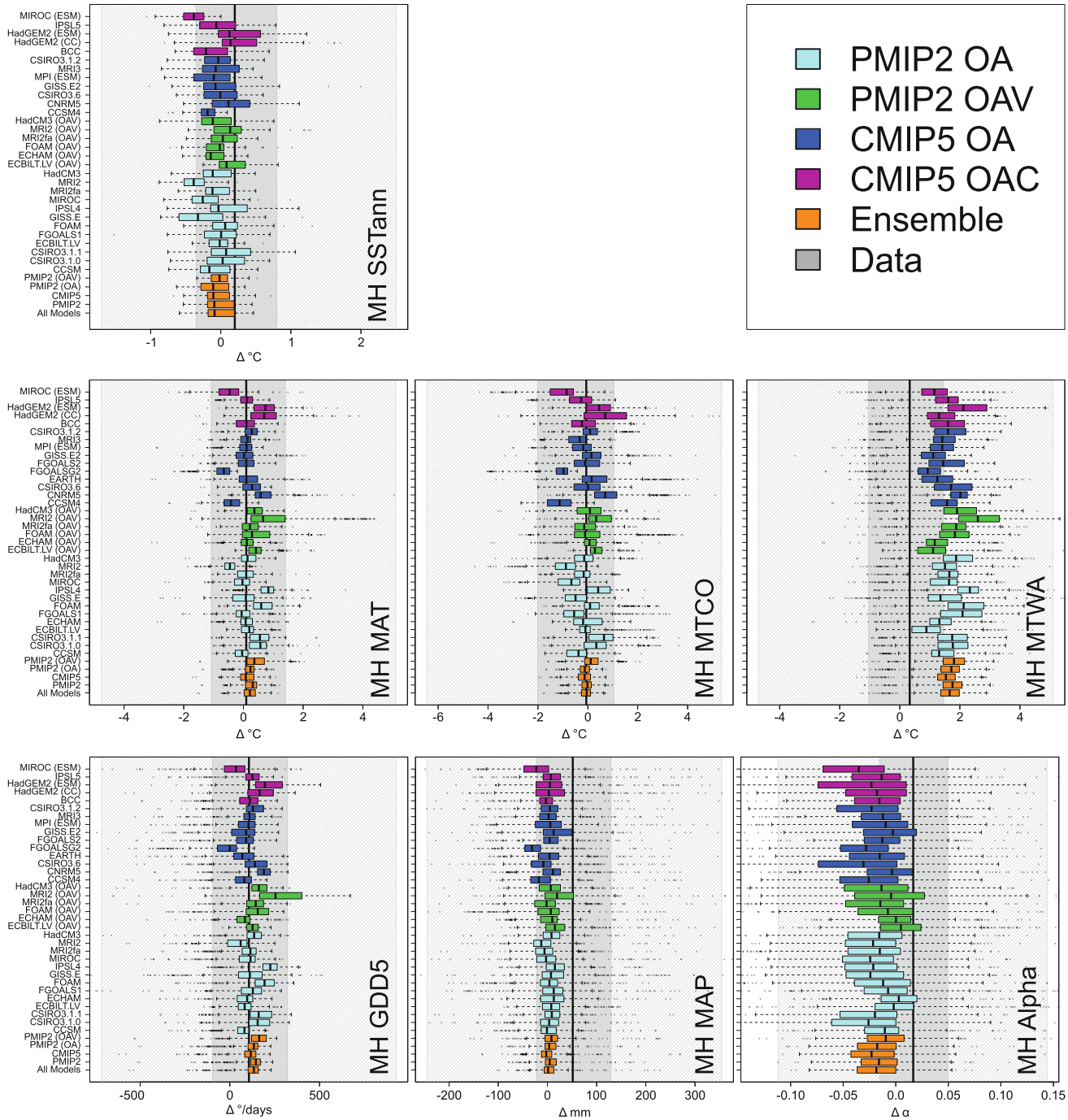


Fig. 6 Comparison of median and interquartile ranges (IQR) of observed and simulated global climate variables at the MH. The comparisons are made using only the model land (or ocean) grid cells where there are observations. Land climates are evaluated against the reconstructed bioclimatic variables (growing degree days above a threshold of 5 °C, *GDD5;* mean temperature of the warmest month, *MTWA;* mean temperature of the coldest month, *MTCO;* mean annual temperature, *MAT;* mean annual precipitation, *MAP;* the ratio of actual to equilibrium evapotranspiration, α). Ocean climates are evaluated against mean annual sea-surface temperature (SSTann) reconstructions. (There are no reconstructions of other ocean variables for the MH). The median value of the observations is shown as a *black vertical line*, the IQR by *dark grey* shading and the 5–95 percentile limits by *light grey* shading. The models are *colour-coded* to show whether they are PMIP2 or CMIP5 simulations, and whether they are ocean–atmosphere (OA), ocean–atmosphere-vegetation (OAV) or OA carbon-cycle (OAC) models. The simulated median for each model is shown by a *vertical line*, the box represents the IQR and the whiskers the 5–95 percentile limits

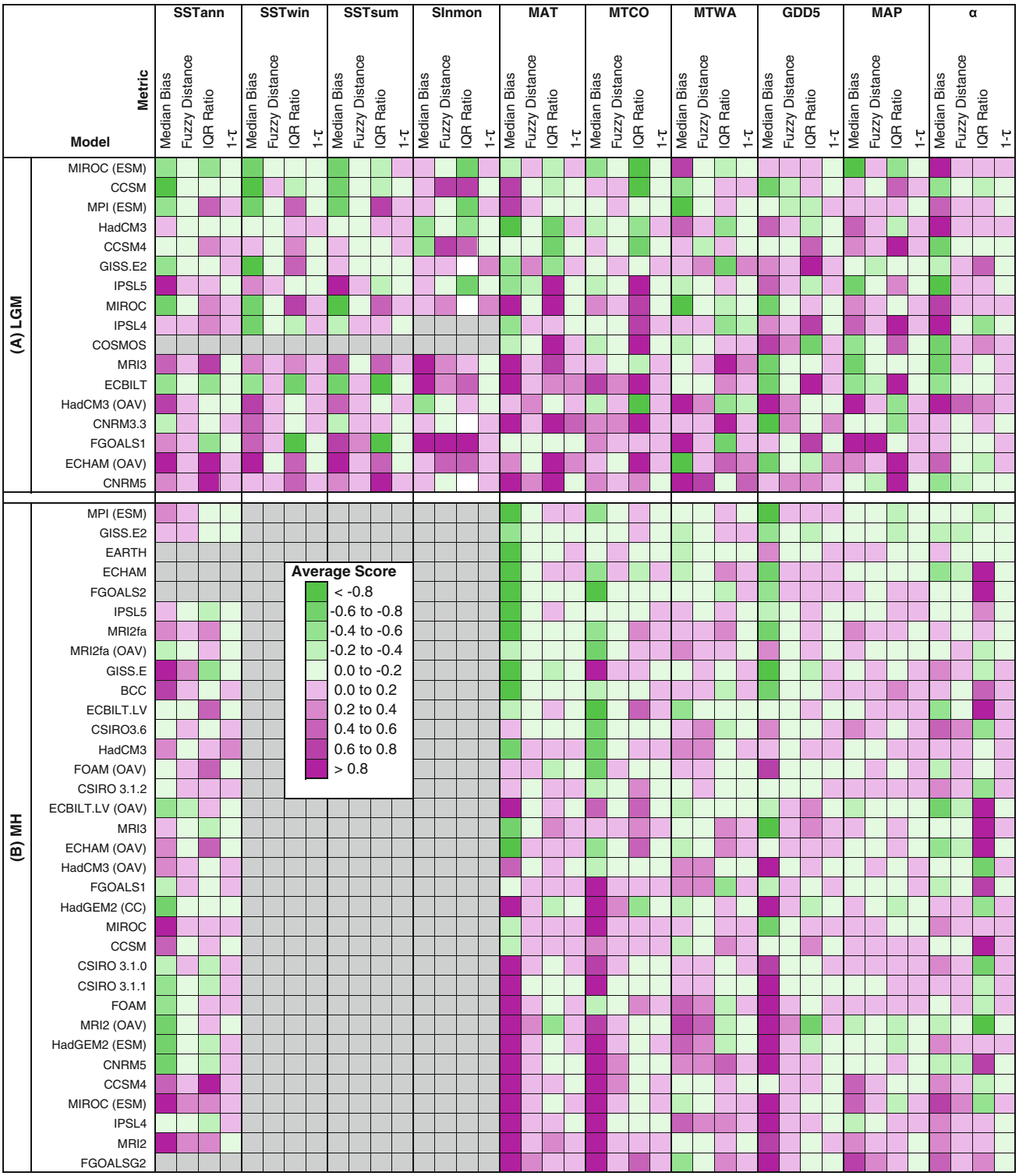


Fig. 7 Summary diagram showing the relative error metrics for (**a**) the Last Glacial Maximum (LGM, ca 21,000 year BP) and (**b**) the mid-Holocene (MH, ca 6,000 year BP) simulations. (Numeric values are given in Tables S9 and S10). Although the number of positive and negative scores must be equal, overall and within each column for each time period, the number of registrations within each positive or negative colour class can differ among variables reflecting the dispersion of the models from the median model. The ordering of the models is based on the average score for the model across all metrics and all variables

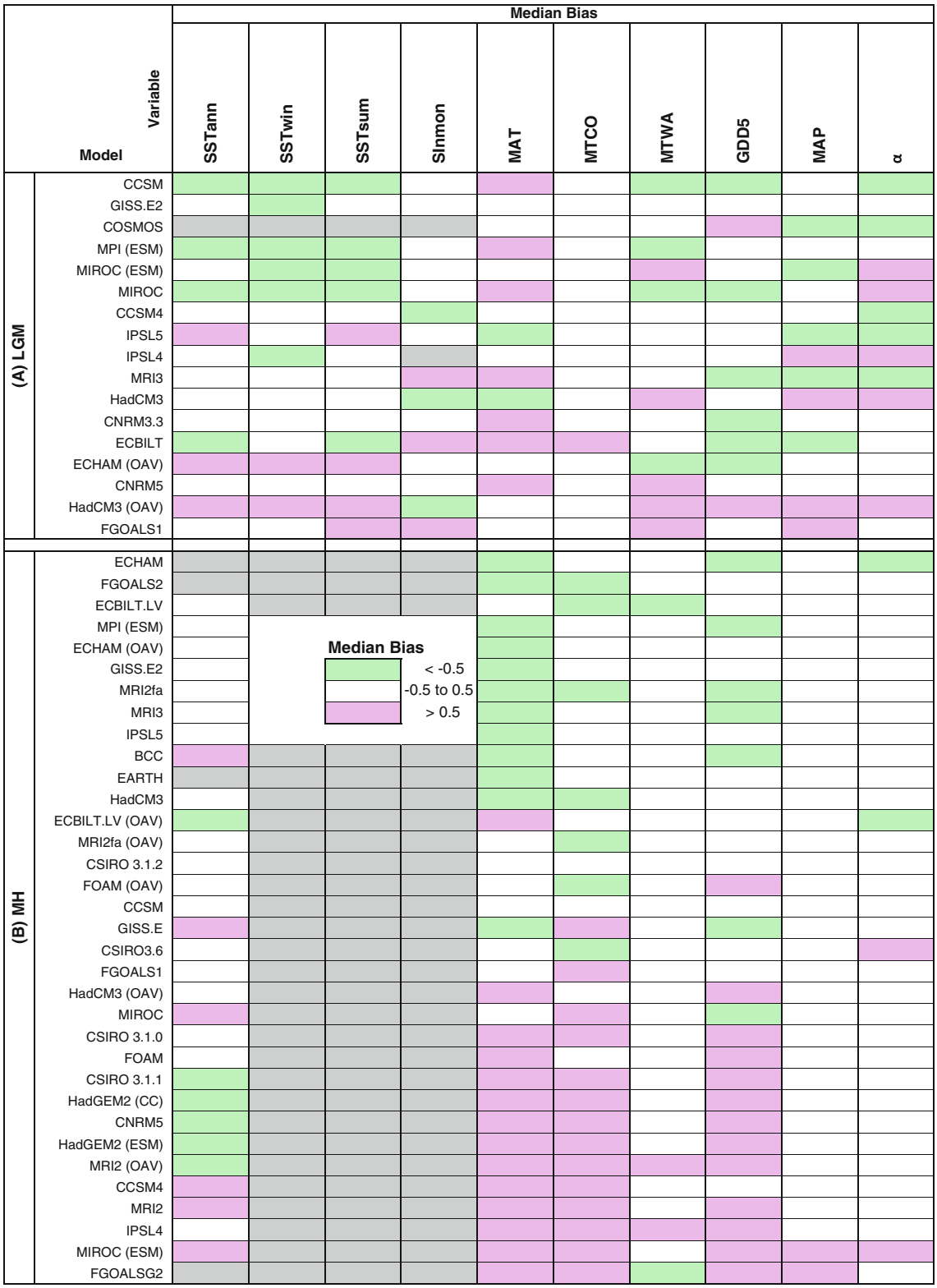


Fig. 8 Summary diagram showing the ranking of models based on the relative error metrics for median bias for all variables at (**a**) the Last Glacial Maximum (LGM, ca 21,000 year BP) and (**b**) the mid-Holocene (MH, ca 6,000 year BP) simulations. Models which score <-0.5 (i.e. a relative error that is 50 % better than the mean model) or >0.5 (i.e. a relative error that is 50 % worse than the mean model) are distinguished. Although the number of positive and negative scores must be equal, overall and within each column for each time period, the number of registrations $>+0.5$ or <-0.5 reflects the dispersion of the models from the median model. Here the models are ordered based on the average score for that model for median bias across all the variables

−0.5 and +0.5) than the median model error on the $1 - \tau$ measure of geographical patterning or fuzzy distance measure (i.e. there is considerable consistency among the models). In contrast, at least some models perform much better (here defined as values <−0.5) than the median model error in terms of biases in the median and in terms of heterogeneity as measured by the IQR. There are also differences in performance between variables: in general, the models are more consistent with one another in their predictions of MTCO than MAT, MTWA or GDD (Fig. 8a). More models perform much better than the median model error for GDD than for either MTWA or MTCO. There are large differences between models in the simulation of MAP and α, but an equal number of models perform much better than the median model error for α and MAP. Over the ocean, there are more models that perform much better than the median model error for SSTs, but the simulation of sea ice cover (SInmon) is much worse (here defined as >0.5) than the median model.

No model performs better than the median model error across all climate variables at the LGM (Figs. 7a, 8a). Nevertheless, the PMIP2 OAV models perform very much worse than the median model error with respect to bias across most variables compared to their OA counterparts (Fig. 7a), as does FGOALS1 and models of the CNRM family (CNRM3.3, CNRM5). These models also perform much worse than the median model error with respect to the IQR metric. Newer versions of a particular model family do not necessarily perform better: CCSM for example performs much better than the median model error for ocean temperature and summer conditions over the land, whereas CCSM4 scores worse than the median model error for these variables. IPSL5 shows an improvement in performance compared to IPSL4 with respect to biases in MAT, MAP and α, but a degradation in the simulation of seasonal and annual SST.

In the MH (Fig. 7b, Table S9), there is more consistency among the models for the $1 - \tau$ and the fuzzy distance metrics (i.e. most models perform only slightly better or worse than the median model error) than for median bias. With the exception of α, there is also considerable consistency between models for the IQR ratio. The simulation of MAP (e.g. as measured by bias: Fig. 8b) is more consistent among models than the simulation of any other climate variable. Although there are models that perform much better than the median model error for MAT, MTCO, MTWA and GDD, there are more models that perform much worse than the median model error for these variables. In contrast, there are more models that perform better than worse than the median model error for SSTann and α. Some models (e.g. ECHAM, ECBILT.LV) consistently perform better than the median model error across all variables as measured by bias, while some models perform better than the median model error for six out of the seven (GISS.E2, ECHAM (OAV), MPI (ESM)) variables. No model consistently performs worse than the median model error across all variables, but some models frequently are very much worse than the median model across several variables (e.g. MIROC (ESM), MRI2, CCSM4, IPSL4, FGOALS2). There is little consistency between the ranking of models with respect to the median model error in the LGM and MH simulations. Furthermore, the ranking of the models depends critically on the choice of metrics and/or variables included (compare Figs. 7 and 8).

3.5 Relationship between LGM biases and climate sensitivity

Reconstructions of LGM climate have been used in attempts to determine the climate sensitivity (see summaries in Edwards et al. 2007; PALAEOSENS Project

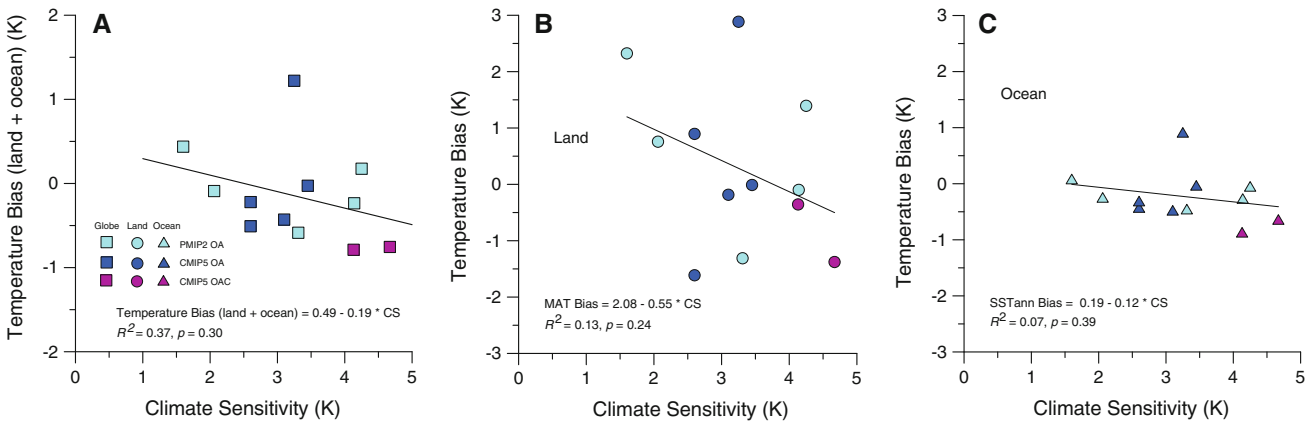


Fig. 9 Comparison of the bias in simulated changes in mean annual temperature at the Last Glacial Maximum (LGM, ca 21,000 year BP) for each model and the climate sensitivity of that model. **a** for global temperature (K) and (**b**) for global land and (**c**) ocean temperature separately. The model biases are only weakly correlated with climate sensitivity

Members 2012). We examine whether the CMIP5/PMIP2 ensemble provides a constraint on climate sensitivity by plotting model temperature bias against climate sensitivity for the PMIP2 (Crucifix 2006; M Crucifix unpublished data) and CMIP5 (Andrews et al. 2012) models, on the assumption that the model that best reproduces reconstructed LGM climates is likely to have the most realistic climate sensitivity under modern conditions. Global biases in simulated LGM temperature are only weakly related to climate sensitivity (Fig. 9); although climate sensitivity (estimated as the point at which the bias in land and ocean temperatures is zero) is ca 2.7 °C, the 12-member model ensemble does not provide a tight constraint. Furthermore, there are opposite biases in ocean and land temperatures: ocean temperatures are globally low and land temperatures globally high, compared with the reconstructions.

4 Discussion

Braconnot et al. (2012) showed that the PMIP2 models reproduce the first-order signals of LGM and MH climate changes. The overall behavior of the CMIP5 models is not different from the PMIP2 models. The models capture major features of past climates such as the differential response of land and ocean to warming/cooling, and the tendency for temperature changes in the higher latitudes to be more extreme than changes in the tropics. Model realism in respect to these first-order signals is important because these signals are features of future projections (Meehl et al. 2007; Sutton et al. 2007; Allan and Soden 2008; Izumi et al. 2013; Li et al. in press). However, although the CMIP5 models have similar success in capturing large-scale climate changes, they display mismatches of similar magnitude between simulated and reconstructed regional climates as their predecessors.

Evaluation of the LGM simulations could provide a constraint on climate sensitivity, by assuming that the model that best reproduces reconstructed LGM climates is likely to have the most realistic climate sensitivity under modern conditions. Using this approach, we estimate a climate sensitivity of 2.7 °C. This is comparable to the estimate of 2.7 ± 0.22 °C made by Annan and Hargreaves (2012) based on the PMIP2 LGM multi-model ensemble and 2.8 °C (with a 90 % confidence range of 1.6–4.7 °C) obtained using an explicitly Bayesian approach with the PMIP2 ensemble by Schmidt et al. (2013). Schmittner et al. (2011), using results from a single model ensemble constrained by glacial MAT anomalies, estimated median climate sensitivity as 2.3 °C with a likely range of 1.4–4.3 °C. The range of climate sensitivities for the CMIP5 (Andrews et al. 2012) and PMIP2 (Crucifix 2006; M Crucifix unpublished data) models is small; the multi-model ensemble used here does not provide a tighter constraint on climate sensitivity than Schmittner et al. (2011). Within this group of models, global biases in simulated LGM temperature are only weakly related to climate sensitivity (Fig. 8). Furthermore, there are opposite biases in ocean and land temperatures: ocean temperatures are globally low and land temperatures globally high, compared with the reconstructions.

Albedo feedback associated with changes in vegetation cover should amplify land cooling (de Noblet et al. 1996; Jahn et al. 2005), particularly in mid- to high-latitudes, and indeed the single pair of PMIP2 simulations which allow us to evaluate the impact of vegetation feedbacks show a substantial additional LGM land cooling (1.2 °C colder than the OA version of the model). However, this simulation also produces (unrealistically) colder oceans than the OA simulation.

The increased atmospheric dust loading at the LGM (Kohfeld and Harrison 2001; Maher et al. 2010), which should contribute to increased cooling, is not included in the LGM experimental design. Model-based estimates (Claquin et al. 2003; Mahowald et al. 2006) show the change in dust forcing is larger over land than over ocean, and the magnitude and even the sign of the forcing varies latitudinally. This could potentially contribute to the latitudinal differences found in the simulated temperature biases (see e.g. Schmittner et al. 2011). Mahowald et al. (2006) showed a small positive forcing over the equatorial oceans, which could help to explain why the present generation of models tends to overestimate SST cooling in the tropics. These two experiments (Claquin et al. 2003; Mahowald et al. 2006) are both constrained by observed LGM dust fluxes but give different results for the magnitude of the global dust forcing (-0.9 and -2.0 Wm^{-2}) and also show different spatial patterns. Furthermore, these dust-forcing estimates do not take account of interactions between dust and clouds. These limitations make it impossible to infer the extent to which inclusion of dust forcing would substantially reduce the biases in simulated LGM temperatures.

The MH simulations also show systematic biases that are different over land and ocean, and between seasons. Land temperature anomalies, particularly in summer, are generally too high and SST anomalies too low. The models underestimate precipitation changes in the regions with the largest summer warming. This bias probably reflects problems in the simulation of land–atmosphere heat fluxes. Wohlfahrt et al. (2004) showed that the IPSL OA model overestimated MH aridity and summer warming in central Eurasia, and this problem was exacerbated by vegetation

feedbacks. The MH PMIP2 OAV simulations (except FOAM) show more warming than the OA version of each model, i.e. the inclusion of interactive vegetation amplifies a problem already present in the OA model.

Previous model evaluations have emphasized regions where models capture the direction of observed climate changes but underestimate those changes (Braconnot et al. 2012). But models can either over- or under-estimate seasonal climate changes. Over-estimation of one variable can be related to underestimation of another, making it possible to infer potential causes of model biases. The discrepancies between simulated and reconstructed climates are generally common to all models. All models overestimate the reconstructed summer cooling of the tropics at the LGM, just as all models underestimate the MH increase of Afro-Asian monsoon precipitation. Explanations of these discrepancies must lie in features common to all models. Nevertheless, some models perform better than others. This is particularly noticeable at the LGM, where the climate-change signal is large, and more consistent across seasons and regions.

Paleoclimate benchmarking provides an independent evaluation of climate models, focusing attention on how well models can simulate climate change. Our results suggest that although models and data are in agreement on the direction and spatial pattern of the large-scale features of climate change (Braconnot et al. 2012; Schmidt et al. 2013; Izumi et al. 2013; Li et al. in press), there are still shortcomings in the amplitude of simulated changes. Recent work by Hargreaves et al. (2013) has shown that this is not a function of the resolution at which the data-model comparisons are made. It is likely that the incorporation of the dust forcing (for the LGM) and improvements to the simulation of vegetation feedbacks (for both LGM and MH) will improve the ability of state-of-the-art models to reproduce past climate changes. However, incorporation of such feedbacks does not obviate the need for continued efforts by modelling groups to achieve accurate simulations of fundamental climate processes.

Acknowledgments We thank our PMIP colleagues who contributed to the production of the benchmark syntheses, and the modelling groups who contributed to the CMIP5 and PMIP2 archives. The analyses and figures are based on data archived at CMIP5 or PMIP2 by March 15th 2012. We thank Hanna Sundquist for her contribution to the compilation of speleothem data, Ann Henderson-Sellers and Gavin Schmidt for their thought-provoking comments, Pascale Braconnot for her persistence in cross-checking the metrics, Sabina Serneels for editorial assistance, and Gavin and Pascale for comments on earlier versions of the text.

References

Allan RP, Soden BJ (2008) Atmospheric warming and the amplification of precipitation extremes. Science 321:1481–1484

Andrews T, Gregory JM, Webb MJ, Taylor KE (2012) Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. Geophys Res Lett 39:L09712. doi:10.1029/2012GL051607

Annan JD, Hargreaves JC (2012) A new global reconstruction of temperature changes at the Last Glacial Maximum. Clim Past Discuss 8:5029–5051. doi:10.5194/cpd-8-5029-2012

Bartlein PJ, Harrison SP, Brewer S, Connor S, Davis BAS, Gajeweski K, Guiot J, Harrison-Prentice TI, Henderson A, Peyron O, Prentice IC, Scholze M, Seppä H, Shuman B, Sugita S, Thompson RS, Viau AE, Williams J, Wu H (2011) Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis. Clim Dyn 37:775–802

Braconnot P, Harrison SP, Otto-Bliesner B, Abe-Ouchi A, Jungclaus J, Peterschmitt J-Y (2011) The palaeoclimate modelling intercomparison project contribution to CMIP5. CLIVAR Exch Newsl 56:15–19

Braconnot P, Harrison SP, Kageyama M, Bartlein PJ, Masson-Delmotte V, Abe-Ouchi A, Otto-Bliesner B, Zhao Y (2012) Evaluation of climate models using palaeoclimatic data. Nat Clim Chang 2:417–424. doi:10.1038/nclimate1456

Chervin RM, Schneider S (1976) On determining the statistical significance of climate experiments with general circulation models. J Atmos Sci 33:405–412

Claquin T, Roelandt C, Kohfeld KE, Harrison SP, Tegen I, Prentice IC, Balkanski Y, Bergametti G, Hansson M, Mahowald N, Rodhe H, Schulz M (2003) Radiative forcing of climate by ice-age atmospheric dust. Clim Dyn 20:193–202

Cramer W, Prentice IC (1988) Simulation of regional soil moisture deficits on a European scale. Norsk Geograf Tids 42:149–151

Crucifix M (2006) Does the Last Glacial Maximum constrain climate sensitivity? Geophys Res Lett 33:L18701. doi:10.1029/2006GL027137

de Noblet N, Prentice IC, Joussaume S, Texier D, Botta A, Haxeltine A (1996) Possible role of atmosphere-biosphere interactions in triggering the last glaciation. Geophys Res Lett 23:3191–3194

Edwards T, Crucifix M, Harrison SP (2007) Using the past to constrain the future: how the palaeorecord can improve estimates of global warming. Prog Phys Geogr 31:481–500

Fairchild IJ, Smith CL, Baker A, Fuller L, Spotl C, Mattey D, McDermott F (2006) Modification and preservation of environmental signals in speleothems. Earth Sci Rev 75:105–153

Giry C, Felis T, Kölling M, Scholz D, Wei W, Lohmann G, Scheffers S (2012) Mid- to late Holocene changes in tropical Atlantic temperature seasonality and interannual to multidecadal variability documented in southern Caribbean corals. Earth Planet Sci Lett 331–332:187–200

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res 113:D06104. doi:10.1029/2007JD008972

Guiot J, Boreux JJ, Braconnot P, Torre F (1999) Data-model comparisons using fuzzy logic in palaeoclimatology. Clim Dyn 15:569–581

Hargreaves JC, Annan JD, Ohgaito R, Paul A, Abe-Ouchi A (2013) Skill and reliability of climate model ensembles at the Last Glacial Maximum and mid-Holocene. Clim Past 9:811–823. doi:10.5194/cp-9-811-2013

Harrison SP, Bartlein PJ (2012) Records from the past, lessons for the future: what the palaeo-record implies about mechanisms of global change. In: Henderson-Sellers A, McGuffie K (eds) The future of the world's climates. Elsevier, Amsterdam, pp 403–436

Izumi K, Bartlein PJ, Harrison SP (2013) Consistent behaviour of the climate system in response to past and future forcing. Geophys Res Lett 40:1–7

Jahn A, Claussen M, Ganopolski A, Brovkin V (2005) Quantifying the effect of vegetation dynamics on the climate of the Last Glacial Maximum. Clim Past 1:1–7

Joussaume S, Taylor KE, Braconnot P, Mitchell JFB, Kutzbach JE, Harrison SP, Prentice IC, Broccoli AJ, Abe-Ouchi A, Bartlein PJ, Bonfils C, Dong B, Guiot J, Herterich K, Hewitt CD, Jolly D, Kim JW, Kislov A, Kitoh A, Loutre MF, Masson V, McAvaney B, McFarlane N, de Noblet N, Peltier WR, Peterschmitt JY, Pollard D, Rind D, Royer JF, Schlesinger ME, Syktus J, Thompson S, Valdes P, Vettoretti G, Webb RS, Wyputta U (1999) Monsoon changes for 6000 years ago: results of 18 simulations from the Palaeoclimate Modelling Intercomparison Project (PMIP). Geophys Res Lett 26:859–862

Kendall M (1938) A new measure of rank correlation. Biometrika 30:81–89

Kohfeld KE, Harrison SP (2001) DIRTMAP: the geological record of dust. Earth Sci Rev 54:81–114

Lachniet MS (2009) Climatic and environmental controls on speleothem oxygen-isotope values. Quat Sci Rev 28:412–432

Leduc G, Schneider R, Kim J-H, Lohmann G (2010) Holocene and Eemian sea surface temperature trends as revealed by alkenone and Mg/Ca paleothermometry. Quat Sci Rev 29:989–1004

Li G, Harrison SP, Bartlein PJ, Izumi K, Prentice IC (in press) Precipitation scaling with temperature in warm and cold climates: an analysis of CMIP5 simulations. Geophys Res Lett 40. doi:10.1002/grl.50730

Maher BA, Prospero JM, Mackie D, Gaiero D, Hesse PP, Balkanski Y (2010) Global connections between aeolian dust, climate and ocean biogeochemistry at the present day and at the last glacial maximum. Earth Sci Rev 99:61–97

Mahowald NM, Muhs DR, Levis S, Rasch PJ, Yoshioka M, Zender CS, Luo C (2006) Change in atmospheric mineral aerosols in response to climate: last glacial period, preindustrial, modern, and doubled carbon dioxide climates. J Geophys Res 111:D10. doi:10.1029/2005JD006653

MARGO Project Members (2009) Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum. Nat Geosci 2:127–132

Meehl GA, Stocker TF, Collins WD, Friedlingstein P, Gaye AT, Gregory JM, Kitoh A, Knutti R, Murphy JM, Noda A, Raper SCB, Watterson IG, Weaver AJ, Zhao Z-C (2007) Global climate projection. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Climate change 2007: the physical science basis. Contribution of working Group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, p 996

Morimoto M, Kayanne H, Abe O, McCulloch MT (2007) Intensified mid-Holocene Asian monsoon recorded in corals from Kikai Island, subtropical northwestern Pacific. Quat Res 67:204–214

Otto-Bliesner BL, Schneider R, Brady EC, Kucera M, Abe-Ouchi A, Bard E, Braconnot P, Crucifix M, Hewitt CD, Kageyama M, Marti O, Rossell-Melé A, Waelbroeck C, Weber S, Weinelt M, Yu Y (2009) A comparison of PMIP2 model simulations and the MARGO proxy reconstruction for tropical sea surface temperatures at last glacial maximum. Clim Dyn 32:799–815

PALAEOSENS Project Members (2012) Making sense of palaeoclimate sensitivity. Nature 491:683–691

Peltier WR (2004) Global glacial isostasy and the surface of the ice-age Earth: the ICE-5G (VM2) model and GRACE. Annu Rev Earth Planet Sci 32:111

Prentice IC, Cramer W, Harrison SP, Leemans R, Monserud RA, Solomon AM (1992) A global biome model based on plant physiology and dominance, soil properties and climate. J Biogeogr 19:117–134

Prentice IC, Sykes MT, Cramer W (1993) A simulation model for the transient effects of climate change on forest landscapes. Ecol Model 65:51–70

Raupach MR (2001) Combination theory and equilibrium evaporation. QJR Meteorol Soc 127:1149–1181

Rencher AC (2002) Methods of multivariate analysis, 2nd edn. John Wiley & Sons, New York

Schmidt GA, Annan JD, Bartlein PJ, Cook BI, Guilyardi E, Hargreaves JC, Harrison SP, Kageyama M, LeGrande AN, Konecky B, Lovejoy S, Mann ME, Masson-Delmotte V, Risi C, Thompson D, Timmermann A, Tremblay L-B, Yiou Y (2013) Using paleo-climate comparisons to constrain future projections in CMIP5. Clim Past Disc 9:775–835. doi:10.5194/cpd-9-775-2013

Schmittner A, Urban NM, Shakun JD, Mahowald NM, Clark PU, Bartlein PJ, Mix AC, Rosell-Melé A (2011) Climate sensitivity estimated from temperature reconstructions of the Last Glacial Maximum. Science. doi:10.1126/science.1203513

Sutton RT, Dong BW, Gregory JM (2007) Land/sea warming ratio in response to climate change: IPCC AR4 model results and comparison with observations. Geophys Res Lett 34:L02701. doi:10.1029/2006gl028164

Taylor KE, Stouffer RJ, Meehl GA (2011) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498. doi:10.1175/BAMS-D-11-00094.1

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498. doi:10.1175/BAMS-D-11-00094

Tran L, Duckstein L (2002) Comparison of fuzzy numbers using a fuzzy distance measure. Fuzzy Sets Syst 130:331–341

Valdes PJ (2011) Built for stability. Nat Geosci 4:414–416

Ventura V, Paciorek CJ, Risbey JS (2004) Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. J Clim 17:4343–4435

Wilks DS (2006) On "field significance" and the false discovery rate. J Appl Meteorol Climatol 45:1181–1189

Wilks DS (2011) Statistical methods in the atmospheric sciences, volume 100, third edition (international geophysics), vol 100. Elsevier Science Publishing Co Inc, Imprint: Academic Press, San Diego, p 676

Wohlfahrt J, Harrison SP, Braconnot P (2004) Synergistic feedbacks between ocean and vegetation on mid- and high-latitude climates during the mid-Holocene. Clim Dyn 22:223–238

Yu KF, Zhao JX, Wei GJ, Cheng XR, Wang PX (2005) Mid-late Holocene monsoon climate retrieved from seasonal Sr/Ca and $d^{18}O$ records of *Porites lutea* corals at Leizhou Peninsula, northern coast of South China Sea. Glob Planet Chang 47:301–316

Zhang Q, Sundqvist HS, Moberg A, Körnich H, Nilsson J, Holmgren K (2010) Climate change between the mid and late Holocene in northern high latitudes—Part 2: model-data comparisons. Clim Past 6:609–626